2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-6819-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIBM55620.2022.9995619

Modeling Annotator Variation and Annotator Preference for Multiple Annotations Medical Image Segmentation

Xutao Guo[1,2], Shang Lu[1], Yanwu Yang[1,2], Pengcheng Shi[1], Chenfei Ye[4], Yang Xiang[2], Ting Ma[1,2,3,4,*]

[1] Electronic & Informatin Engineering School, Harbin Institute of Technology (Shenzhen), Shenzhen, China

[2] Peng Cheng Laboratory, Shenzhen, China

*[3] Guangdong Provincial Key Laboratory of Aerospace Communication and Networking Technology,
Harbin Institute of Technology (Shenzhen), Shenzhen, China 518055*

[4] International Research Institute for Artifcial Intelligence, Harbin Institute of Technology (Shenzhen), Shenzhen, China

[*]Corresponding author: tma@hit.edu.cn

***Abstract*—Medical image segmentation annotation suffers from annotator variation due to the inherent differences in annotators' expertise and the inherent blurriness of medical images. In practice, using opinions from multiple annotators can effectively reduce the impact of such annotator-related biases. Meanwhile, it is common practice in deep learning to fuse multiple annotations through methods such as majority voting, but these methods ignore the rich information of annotator preferences ingrained in the original multi-annotator annotations. To address this issue, we propose a modeling annotator variation and annotator preference (AVAP) framework for multiple annotations medical image segmentation, which consists of three parts. First, the widely used encoder-decoder backbone network use to extract feature maps of the image. Second, an annotator variation modeling (AVM) module is devised to estimate the annotation variation among multiple annotators by modeling multi-annotations as a multi-class segmentation problem. Third, an annotator preference modeling (APM) module estimate each annotator's preference-involved segmentation by annotator encoding and dynamic filter learning. The experiment on the RIGA benchmark with multiple annotations shows that our AVAP framework outperforms a range of state-of-the-art (SOTA) multiple annotations segmentation methods. Further, we are the first to introduce dynamic filter learning into the annotator preference modeling.**

***Index Terms*—medical image segmentation, multi-annotations, annotator preference, dynamic filter learning**

I. Introduction

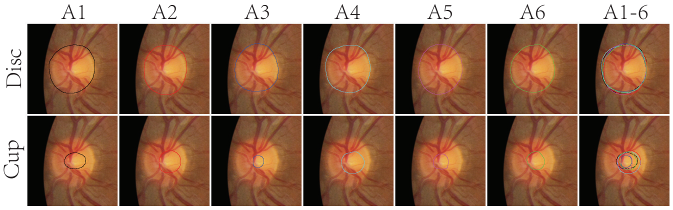

Fig. 1: An example of medical image annotation by six annotators (A1-6) on the RIGA benchmark.

TABLE I: Examining the grading consistency of individual annotators on the RIGA test set (measured by Dice (D_{disc} (%), D_{cup}(%))). M1-6 denote Res-U-Net supervised by individual annotator's annotation. The A1-6 indicates the predictions of each model are evaluated against each annotator's annotation.

	A1	A2	A3	A4	A5	A6
M1	**95.9,84.4**	94.8,81.2	95.1,79.5	95.9,79.1	95.6,79.4	95.9,75.8
M2	95.3,84.0	**96.1,84.7**	96.1,80.8	96.1,81.8	96.5,80.3	96.3,77.4
M3	95.4,82.5	94.9,81.1	**96.8,83.6**	95.8,80.3	96.3,81.1	96.2,76.3
M4	95.1,80.3	95.6,82.1	96.3,77.4	**96.4,87.9**	96.1,72.7	96.4,68.7
M5	95.1,83.6	94.9,80.0	96.0,81.9	96.3,75.5	**96.8,84.0**	96.1,79.4
M6	95.5,81.4	95.6,80.0	96.3,78.9	96.2,74.5	96.4,82.3	**97.1,80.2**

Segmentation and quantitative evaluation of regions of interest in medical images are of great importance in formulating therapeutic strategies, monitoring the disease's progress, and predicting the prognosis of patients [1], [2]. Data-driven methods such as deep convolution neural networks (DCNN) have recently achieved state-of-the-art performance in medical image segmentation [3]–[5]. As we all know, one of the basic facts contributing to this success relies on supervised learning of dense pixel annotations of images. Despite the success of the aforementioned CNN-based methods, medical image segmentation annotation suffers from annotator variation due to the inherent differences in annotators' expertise and the inherent blurriness of medical images [6]. For example, Suetens et al. [7] showed that the three trained annotators (two radiologists and one radiotherapist) delineated a lesion of the liver in an abdominal CT image twice with an interval of about one week, resulting in the variation of delineated areas up to 10% per observer and more than 20% between observers.

In practice, we tend to generate multiple plausible hypotheses when faced a situation that we are not sure about. Especially in practical clinical applications, medical images are often annotated by multiple experts to mitigate the subjective bias of a particular expert due to factors such as level of expertise or possible omission of subtle symptoms [10], [11]. It is worth noting that the annotations provided by multiple annotators are reasonable, albeit with different preferences [9]. For example,

Xutao Guo and Shang Lu are co-first authors.

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on March 24,2023 at 06:50:19 UTC from IEEE Xplore. Restrictions apply.

annotators who advocate aggressive treatment often delineate a slightly larger lesion area than the area marked by others. To quantitatively and qualitatively demonstrate the annotator's preference, one preliminary experiments are conducted on the RIGA benchmark [16]. And six glaucoma experts from different organizations labeled the optic cup and disc contour masks manually for the RIGA benchmark. We train the Res-U-Net using individual annotator's annotations, and obtain six different models (named M1-6) corresponding to six annotators (named A1-6). Table I lists the Dice coefficient of each model for six different annotations. It is obvious that all the models have the optimal performance when trained and evaluated with the same annotator's annotations but much worse when evaluated by others' annotations. Therefore, a single annotator has a specific and consistent scoring pattern, called annotator preference [28].

It is a common practice in deep learning to fuse multiple annotations (such as majority voting [17], label fusion [18], [20], [34], or label sampling [19]) to reduce multiple annotation variation, but these methods overlook the rich information of annotators' preferences ingrained in the raw multi-annotator annotations. Particularly, converting the multiple annotations of each training image into a proxy ground truth may lead the segmentation result to be neither fish nor fowl [30]. Currently, there are few studies related to annotator preference modeling in multiple annotations medical image segmentation. The most direct method is to train multiple base models for the annotations of a single annotator, but this method is more complicated and does not jointly utilize the information of multiple annotations. In addition, the complexity of existing models based on multi-head network structure [17], [30] will increase with the number of annotators.

To this end, we propose a framework for modeling annotator variation and annotator preference (AVAP) for multiple annotations medical image segmentation. Unlike the fuse method that reduces the impact of annotation variation, AVAP instead advocates modeling annotator preferences separately from annotator variation. So AVAP not only produces calibrated image segmentation (output of AVM), but also mimics each annotator and segments medical images with annotator preference (output of APM). AVAP consists of three parts: First, the widely used encoder-decoder backbone is used to extract feature maps of the image. The extracted feature maps are used for subsequent annotator preference and annotator variation modeling. Second, an annotator variation modeling (AVM) module is devised to estimated the annotation variation among multiple annotators by modeling multi-annotations as multi-class segmentation problem. It should be noted that we achieved the first best result on the MICCAI QUBIQ 2021 challenge [8], [21] leaderboards by modeling multi-annotation variation as a multi-class segmentation problem. Third, an annotator preference modeling (APM) module estimate each annotator's preference-involved segmentation by annotator encoding and dynamic filter learning [22]. The convolution kernel parameters in the dynamic network are adaptively generated based on the input feature maps and annotator coding. Specifically, the annotator specific prior is fed to the controller to guide the generation of dynamic network convolution kernel parameters for each annotator. In this design, compared with the multi-head network, the complexity caused by the number of annotators can be ignored. The main contributions of this paper are as follows:

1) We propose a modeling annotator variation and annotator preference (AVAP) framework for multiple annotations medical image segmentation. AVAP not only produces calibrated image segmentation (output of AVM), but also mimics each annotator and segments medical images with annotator preference (output of APM).
2) Different from existing state-of-the-art methods, AVAP adopts a completely different network structure and makes a more novel exploration. AVAP estimated the annotation variation by modeling multi-annotations as a multi-class segmentation problem. We used this method to achieve the first best result on the MICCAI QUBIQ 2021 challenge leaderboards. Further, we are the first to introduce dynamic filter learning into the annotator preference modeling.
3) The experiment on the RIGA benchmark shows AVAP outperforms a range of state-of-the-art multiple annotations medical image segmentation methods.

II. RELATED WORK

A. Medical Image Segmentation

With the advancement of CNNs, an increasing number of deep learning architectures have been proposed for medical image segmentation [12], [13]. Particularly, U-Net [4] is one of the most commonly used convolutional network structures in medical image segmentation. By adopting an encoder-decoder network structure and skip connection, it can combine features of the different decoding layers with features of the different encoding layers. Oktay et al. [23] introduced the attention mechanism [14] into U-Net, which can suppress irrelevant areas in the input image and highlight the salient features of specific local areas. Gu et al. [24] integrated dense atrous convolution block [15] and residual multi-kernel pooling to U-Net structure to capture high-level features with context information. Zhou et al. [25] proposed a new segmentation architecture based on nested and dense skip connections to reduces the gap between the feature maps of the encoding and decoding sub-networks. A common practice adopted by the above methods typically requires unique ground-truth annotations, each pairing with one of the input images to train the deep learning models.

B. Medical Image Segmentation with Multiple Annotations

A few methods have been proposed to address the issue of multiple annotations in medical image segmentation. Here we summarize these methods roughly into two categories, as follows:

1) Fusing multiple annotations: This kind of method is a common practice to take majority voting [17], label fusion [18], [20], [34], or label sampling [19] to construct training examples by retaining unique ground-truth labels for each of the training instances. Jensen et al. [19] propose a better calibrated model is

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on March 24,2023 at 06:50:19 UTC from IEEE Xplore. Restrictions apply.

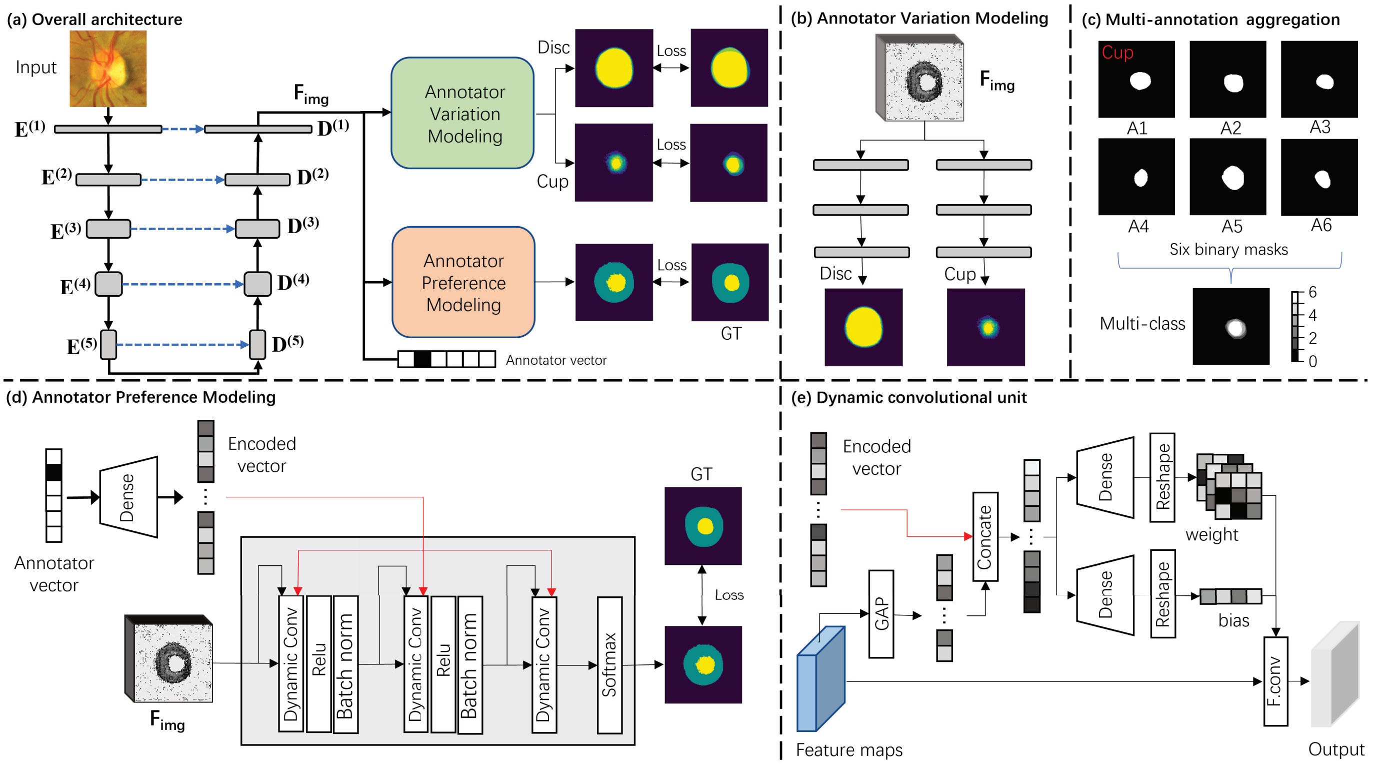

Fig. 2: Architecture of AVAP framework, which consists of three parts. First, the backbone network is used to extract feature maps of the image. Second, an annotator variation modeling (AVM) module is devised to estimate the annotation variation among multiple annotators by modeling multi-annotations as multi-class segmentation problem. Third, an annotator preference modeling (APM) module estimate each annotator's preference-involved segmentation by dynamic filter learning.

obtained when training with a label sampling scheme that takes advantage of inter-annotator variability during training. Guan et al. [17] predicted the gradings of each annotator individually and learned the corresponding weights for final prediction. Mirikharaji et al. [29] propose a spatially adaptive reweighting approach to treat multiple noisy pixel-level annotations commensurately in the loss function. Although these strategies are simple and easy to implement, they completely ignore the information of the preference of annotators.

2) Modeling annotator preference: Currently, there are few studies related to annotator preference modeling in multiple annotations medical image segmentation. The most direct method is to train multiple base models for the annotations of a single annotator, but this method is more complicated and does not jointly utilize the information of multiple annotations. Ji et al. [28] proposed MRNet framework to incorporate the multi-rater (dis-)agreement cues and generate calibrated model predictions that better reflected the underlying agreement among multiple experts. Liao et al. [30] propose the PADL framework based on multi-head network structure, which treats the annotation bias as the combination of annotator's preference and stochastic errors. However, the complexity of multiple-head/branch [17], [31] architecture will increase with the number of annotators. Different from above works, AVAP adopts a completely different network structure and makes a more novel exploration.

C. Dynamic Filter Learning

In a traditional convolutional layer, the learned filters stay fixed after training. In contrast, the filters of dynamic filter learning are generated dynamically conditioned on an input [26], [32]. Chen et al. [22] dynamically aggregate multiple convolution kernels according to the attention of each input via dynamic convolution to improve the representation ability. Zhang et al. [33] uses dynamic filter learning to address the partially labelling issue for multi-organ and tumor segmentation. Tian et al. [35] propose a simple yet effective instance segmentation framework by using instance-wise ROIs as inputs to a network of fixed weights. The above methods show that dynamic filter learning can increase the flexibility of the network and enhance the representation ability. In this paper, we employ the dynamic filter learning to model annotator preference. And the convolution kernel parameters in the dynamic network are adaptively generated based on the input feature maps and annotator coding.

III. Method

A. Overall Framework

Let N medical images annotated by M annotators be denoted by $D = \{x_i, y_i^1, y_i^2, ..., y_i^M\}_{i=1}^N$. x_i represents the i-th image, and y_i^m is the annotation given by the m-th annotator. Our

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on March 24,2023 at 06:50:19 UTC from IEEE Xplore. Restrictions apply.

goal is to train a segmentation model on the multi-annotations data set so that the model can generate calibrated segmentation [27] and mimic each annotator's preference.

Figure 2 (a) shows that the AVAP framework consists of an encoder-decoder backbone, an AVM model, and an APM model. First, the backbone network extracts the feature maps F_{img} of the input image. The extracted feature maps are used for subsequent AVM models and APM models. Based on F_{img}, the AVM model estimates a calibrated segmentation map, which approximates the mean voting of M annotations. AVM model multi-annotations as multi-class segmentation problem. We evaluate the calibrated predictions of the AVM using the soft dice coefficient metric. Figure 2 (c) shows how to convert multiple binary annotations into single multi-class masks. Based on F_{img}, the APM module estimate each annotator's preference-involved segmentation by annotator encoding and dynamic filter learning. The convolution kernel parameters in the dynamic network are adaptively generated by the controller according to the input feature maps and annotator specific prior coding.

B. Encoder-decoder Backbone

AVAP adopts the widely used U-Net architecture with a ResNet34 [37] pre-training from ImageNet [38] as the encoder part (Corresponding to the eight coding layers of the ResNet34). Symmetrically, the decoder consists of five blocks that progressively upsample the feature maps to restore their resolution. The skip connections were also employed between down-convolutional layers and up-convolutional layers. In each of the first four blocks, the feature maps are upsampled by a transposed convolutional layer with stride 2. Then, the feature maps are processed by a convolutional layer, concatenated with feature maps from the encoder, and fed to a ReLU layer and a batch normalization layer. The last decoder block only upsamples the image features to the original image size using a transposed convolutional layer with stride 2 [30]. As a result, the backbone network generates feature maps F_{img} of the same size for each input image, shown as follows, which contain rich semantic information.

$$F_{img} = F_B(x;\theta_B), \tag{1}$$

where θ_B represent the parameters in the backbone network F_B.

C. Annotator Variation Modeling (AVM) Module

The AVM module is devised to estimate the calibrated predictions among different annotators by modeling multi-annotations as a multi-class segmentation problem. As shown in Figure 2 (b), AVM contains two branches, corresponding to the Cup and Disc in the RIGA dataset, respectively. Each branch contains three blocks. In each of the first two blocks, two convolutional layers are repeatedly employed, each followed by a ReLU layer and a batch normalization layer. At the third blocks, one 1×1 convolutional layers is used to map feature maps to seven classes.

Figure 2 (c) shows how to convert multiple binary annotations into single multi-class masks. Taking the RIGA benchmark as an example, the six masks will be converted to two multi-classes (including seven class) masks. This method is simple and yields well-calibrated predictions. Taking the cup category as an example, Figure 2 (c) shows the aggregation of six binary masks into a single multi-class mask $M(x_i)$ as

$$M(x_i) = \sum_{m=1}^{M} y_i^m, \quad 0 \leqslant M(x_i) \leqslant M \tag{2}$$

where M denotes the number of annotators. y_i is the binary annotation of x_i annotated by i-th annotator. Therefore, the value of the aggregated multi-class mask M is between 0 and the number of annotators. The value of each pixel of M represents the number of annotators for which it is marked as the target pixel. By aggregating multiple annotation binary masks, we modeling multi-annotations as multi-class segmentation problem with $M+1$ class (including background). The class m marks the agreement of exactly m annotators, for $m \in \{0, 1, ..., M\}$. Figure 2 (b) shows the two branches outputs the calibrated predictions based on F_{img} as

$$\begin{aligned} P_{disc} &= F_{disc}(F_{img};\theta_{disc}) \\ P_{cup} &= F_{cup}(F_{img};\theta_{cup}) \end{aligned} \tag{3}$$

where θ_{disc} and θ_{cup} represent the parameters in the branche F_{disc} and branche F_{cup}, respectively. The output mask P thus represents the pixels that would be marked by M annotators. Further, dividing the output mask values with the number of annotators M results in values on the interval $[0, 1]$, which can also be interpreted as annotation or segmentation (un)certainty and reflects the uncertainty of the expert annotators [36].

D. Annotator Preference Modeling (APM) Module

1) Annotator encoding: Taking the RIGA benchmark as an example, each annotation corresponds to a fixed annotator. This information is a critical prior that tells the model with which annotator it is dealing and what kind of preferred segmentation results should it produce. For instance, given an input image, AVAP is expected to be specialized for a certain annotator, i.e., predicting the segmentation result with this annotator's preference. In the APM module, the expertise level cues of multiple annotators are formed as a normalized annotators vector $V \in \mathbb{R}^{1\times1\times M}$, shown as follows.

$$V_m = \begin{cases} 0 & if \quad m \neq i \\ 1 & otherwise \end{cases} \qquad m = 1, 2, ..., M \tag{4}$$

Where M represents the total number of annotators. i corresponds to a specific annotator. It is fed to the APM module for annotator-awareness. $V_m = 1$ means that the annotation of m-th annotation is available for the current input x [33].

2) Dynamic filter: In a traditional convolutional layer, the learned filters stay fixed after training. Therefore, a network optimized for one task is necessarily less optimized for other tasks. So it is difficult to use a single network to perform the task of preference segmentation for multiple annotators. In contrast, the filters of dynamic filter learning are generated

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on March 24,2023 at 06:50:19 UTC from IEEE Xplore. Restrictions apply.

dynamically conditioned on an input [32]. In this paper, we introduce dynamic filter learning to modeling annotator preference, which is specialized to segment a particular annotator. As in the paper [33], a single convolutional layer is used as a task-specific controller $\varphi()$. The input feature maps F is aggregated via global average pooling (GAP) and concatenated with the annotator encoding vector V_m as the input of $\varphi()$. Then, the kernel parameters w are generated dynamically conditioned not only on the assigned annotator encoding V_m but also on the input feature maps F itself, expressed as follows

$$w^m = \varphi(GAP(F)||V_m; \theta_\varphi) \tag{5}$$

where θ_φ represents the controller parameters, and $||$ represents the concatenation operation.

Figure 2 (d) shows the APM contains three dynamic convolutional layers. The kernel parameters in three layers, denoted by $\omega_k^m = \{\omega_1^m, \omega_2^m, \omega_3^m\}$, are dynamically generated by the controller $\varphi()$ according to the input feature maps F and annotator encoding vector V_m. In each of the first two dynamic convolutional layers is with $3 \times 3 \times 3$ kernels. The last dynamic convolutional layer is with $1 \times 1 \times 1$ kernels. The annotator's preference predictions of image with regard to m-th annotator is computed as

$$P_m = BN(Relu(BN(Relu(F_{img} * \omega_1^m)) * \omega_2^m)) * \omega_3^m \tag{6}$$

where $*$ represents the convolution, and P_m represents the predictions of m-th annotator's preference.

E. Training and Testing

1) Training: After comparative experiments, this paper adopts a special training strategy, which is divided into two steps. First, we jointly train the backbone network and AVM module. Then the backbone network and AVM module are frozen and the APM module is trained. Furthermore, during APM training, we randomly select different annotators' codes for training. We jointly use the Dice loss and Cross-entropy loss as the objective for two steps. The loss function L_{total} is formulated as

$$L_{total} = L_{Dice} + L_{CE} \tag{7}$$

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{I} p_i y_i}{\sum_{i=1}^{I} p_i + y_i + \epsilon} \tag{8}$$

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{K} y_{ic} log(p_{ic}) \tag{9}$$

where p_i and y_i represent the prediction and ground truth of i-th voxel, N is the number of all voxels. K is the number of categories.

2) Testing: During inference, the proposed AVAP is flexible to m segmentation annotations. Given a test image, the feature maps F_{img} is extracted from the backbone network. Based on F_{img}, the AVM model estimates a calibrated segmentation, which approximates the mean voting of M annotations. Based on F_{img}, the APM module estimate each annotator's preference-involved segmentation by annotator encoding and dynamic filter learning. In addition, if annotations are all required, AVAP is able to efficiently segment all of m annotations in turn.

IV. Experiments and Results

A. Datasets

RIGA: The RIGA benchmark [21] is a publicly available dataset for retinal cup and disc segmentation, which contains in total of 750 color fundus images from three sources, including 460 images from MESSIDOR, 195 images from BinRushed and 95 images from Magrabia. Six glaucoma experts from different organizations labeled the optic cup/disc contours manually in each image. We followed the data split scheme used in [28], [30], using 655 samples from BinRushed and MESSIDOR for training and 95 samples from Magrabia for test.

B. Experimental Setup

1) Implementation Details: In this study, all the networks train using Pytorch using NVIDIA TESLA V-100 (Pascal) GPUs with 32 GB memory. All images were normalized via subtracting the mean and dividing by the standard deviation on a pixel-by-pixel basis. The mean and standard deviation were counted on training cases. We optimized our methods with the Adam optimizer with the learning rate 1e-4 and the weight decay 1e-5. The batch size is set to 8. And all training and test images are uniformly resized to the dimension of 256×256 pixels.

2) Evaluation Metric: The output of the AVM is to produce probability map that can reflect the underlying inter-rater agreement/disagreement, i.e., calibrated predictions, for medical image segmentation. In order to better evaluate the calibrated model predictions, we use soft dice coefficient metric through multiple threshold levels, set as (0.1, 0.3, 0.5, 0.7, 0.9) in this paper, instead of using a single threshold (e.g., 0.5) [28]. At each threshold level, the threshold is applied to the predicted probability map and mean voting of annotations to generate hard Dice. The Dice scores obtained at multiple thresholds are averaged and then we obtain the soft metrics, denoted as D^s. Based on Soft Dice, there are two performance metrics, namely 'Average' and 'Mean Voting'. 'Mean Voting' is the Soft Dice between the predicted calibrated segmentation and the mean voting annotations. Higher 'Mean Voting' represents better performance on modeling calibrated segmentation. The annotator-specific predictions are evaluated against each annotator's delineations, and the average Soft Dice is denoted as 'Average'. Higher 'Average' represents better performance on miming each annotator [30].

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on March 24,2023 at 06:50:19 UTC from IEEE Xplore. Restrictions apply.

TABLE II: Quantitative results with different methods on the RIGA test set. Here, we use soft Dice metrics (D^s_{disc} (%), $D^s_{cup}(\%)$) to evaluate calibrated segmenttaion (corresponding to 'Mean Voting'). The predictions of each model are evaluated against each annotator's annotation (D_{disc} (%), $D_{cup}(\%)$) and the average performance for the six annotations (corresponding to 'Average') is also given. The best results in each column are highlighted.

Methods	A1	A2	A3	A4	A5	A6	Average	Mean Voting
M1	95.93, 84.39	94.76, 81.15	95.06, 79.52	95.90, 79.05	95.62, 79.40	95.96, 75.80	95.56, 79.89	96.02, 81.72
M2	95.32, 84.02	96.06, 84.67	96.13, 80.79	96.14, 81.79	96.51, 80.33	96.32, 77.39	96.08, 81.49	95.80, 82.42
M3	95.43, 82.52	94.86, 81.09	96.79, 83.55	95.82, 80.28	96.27, 81.07	96.19, 76.31	95.89, 80.80	95.36, 81.20
M4	95.14, 80.31	95.63, 82.08	96.33, 77.42	96.42, 87.89	96.10, 72.70	96.42, 68.69	96.01, 78.18	96.11, 79.24
M5	95.06, 83.62	94.92, 79.99	96.00, 81.88	96.27, 75.47	96.75, 83.97	96.07, 79.40	95.85, 80.72	95.88, 80.25
M6	95.50, 81.39	95.64, 80.00	96.25, 78.92	96.19, 74.47	96.38, 82.32	97.09, 80.22	96.18, 79.55	96.03, 79.63
MH-UNet	96.03, 85.30	95.98, **85.69**	96.90, 83.99	96.68, 84.86	**97.12**, 83.06	96.78, 77.48	96.58, 83.40	96.91, 84.35
MV-UNet	95.06, 84.33	95.27, 82.57	96.05, 79.35	95.48, 80.29	96.26, 81.05	95.33, 78.11	95.57, 80.95	97.35, 85.74
LS-UNet	95.25, 83.43	94.71, 80.10	95.92, 81.41	96.30, 78.57	96.13, 82.19	96.04, 79.15	95.73, 80.81	97.21, 81.37
CM-Net [30]	96.29, 84.59	95.46, 81.44	96.60, 81.84	96.90, 87.52	96.86, 82.39	96.93, 78.82	96.51, 82.77	96.64, 81.96
MR-Net [28]	95.35, 81.77	94.81, 81.18	95.80, 79.23	95.96, 84.46	95.90, 79.04	95.76, 76.20	95.60, 80.31	97.55, 87.20
PADL [30]	96.40, 85.22	95.60, 85.15	96.64, 82.76	96.82, 88.79	96.78, 83.45	96.87, 79.72	96.52, 84.18	97.65, 87.75
AVAP	**96.40, 85.66**	**96.24**, 85.61	**96.97, 84.21**	**97.12, 89.00**	96.92, **84.15**	**97.08, 82.15**	**96.78, 85.13**	**97.88, 87.90**

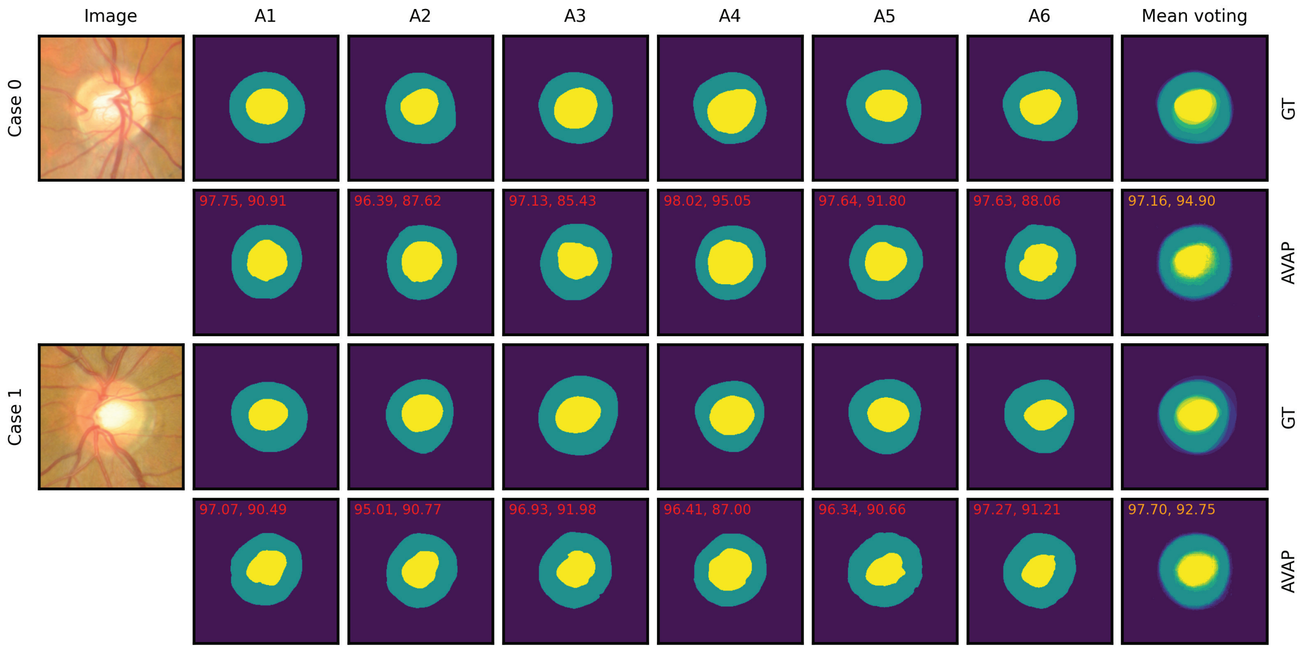

Fig. 3: Visualization of calibrated segmentation maps predicted by AVM and six annotator preference segmentations by APM. $A1 - 6$ corresponds to the annotations of six different annotators. The top left corner of the prediction results in red shows the hard Dice (D_{disc} (%), $D_{cup}(\%)$) against each annotator's annotation. The top left corner of the prediction results in orange shows the soft Dice (D^s_{disc} (%), $D^s_{cup}(\%)$) against mean voting annotation.

C. Comparison Results

We conduct quantitative experiments to compare our AVAP with a range of multiple annotations segmentation methods on the RIGA test set in Table II. Here, M1-M6 refer to the Res-U-Net base model trained with the corresponding labels graded by annotators 1-6 (A1-6), respectively; A variant of Res-U-Net with multiple segmentation heads, each used to mimic annotations from a specific annotator defined as MH-UNet; A Res-U-Net trained with the mean voting of annotations is defined as MV-UNet; A Res-U-Net trained with randomly selected annotation from the candidate annotations of each sample is defined as LS-UNet [19]; CM-Net is an annotator bias disentangling method that uses a confusion matrix to model human errors [39]. MR-Net [28] and PADL [30] are existing state-of-the-art methods for modeling the multi-annotators (dis-)agreement.

Table II shows that $M1 - 6$ have the optimal performance when trained and evaluated with the same annotator's annotations but much worse when evaluated by others' annotations. The AVAP consistently achieves superior performance under most conditions, reflecting that AVAP not only produces calibrated image segmentation but also mimics each annotator and segment medical images with annotator preference. Meanwhile, since the annotator-related bias is considered, MR-Net, MH-UNet, and PADL perform well no matter being evaluated

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on March 24,2023 at 06:50:19 UTC from IEEE Xplore. Restrictions apply.

TABLE III: Ablation study on the effect of pre-training on ImageNet.

Model	Average	Mean Voting
AVAP w/o ImageNet pre-training	96.31, 82.97	97.64, 85.16
AVAP w ImageNet pre-training	96.78, 85.13	97.88, 87.90

TABLE IV: Ablation study on the effect of different conditions (input feature maps, annotator encoding) during the dynamic filter generation.

Task coding	Feature coding	Average	Mean Voting
√	×	96.58, 84.30	97.88, 87.83
×	√	96.18, 82.34	97.88, 87.83
√	√	96.78, 85.13	97.88, 87.90

against the mean voting or individual annotation.

Compared with MR-Net and PADL state-of-the-art methods, AVAP has a good improvement in modeling annotator preference on cup segmentation. In addition, the accuracy of the disc has reached a high accuracy, even the base model (M1-6) can achieve high accuracy. So the corresponding improvement is relatively small. It can also be seen from Figure 1 that the annotations difference between multiple annotators of the disc is small. And AVAP obtains the best-calibrated segmentation results on both disc and cup. Compared with MR-Net and PADL state-of-the-art (SOTA) methods, AVM only adopts a very simple network structure. This further illustrates the effectiveness of modeling multi-annotations as a multi-class segmentation problem. In future work, we can extend the AVM module to introduce better structures or loss functions to further improve the performance of AVM.

We visualized the segmentation maps predicted by AVAP of two cases from the RIGA dataset in Figure 3. It shows that AVM can produce more accurate calibrated segmentation. And the APM can generate segmentation results with different annotator preferences.

D. Ablation study

In this section, we conducted ablation studies on the RIGA dataset to investigate the effectiveness of the detailed design of the AVAP. We use the ‘Average’ and ‘Mean Voting’ as two evaluation indicators for a fair comparison.

1) Pre-training On ImageNet: In deep learning, feature representations learned on a pre-training task can transfer useful information to target tasks [40]. And AVAP adopts the widely used U-Net architecture with a ResNet34 pre-training from ImageNet as the encoder part (Corresponding to the eight coding layers of the network). To evaluate the contributions of pre-training on ImageNet, we compared the proposed AVAP framework with its variants that not use pretrained from ImageNet, i.e., ‘AVAP w ImageNet pre-training’ and ‘AVAP w/o ImageNet pre-training’. As shown in Table III, the ‘AVAP w ImageNet pre-training’ methods achieves the best results. And pre-training on ImageNet has a greater impact on cup segmentation and less impact on disc.

2) Annotator coding & Input Feature Maps in APM: The dynamic filters generated in AVAP are conditioned not only on the input feature maps but also on the assigned annotator. Here we analyze the effect of input feature maps and annotator encoding on model performance. As shown in Table IV, the annotator encoding plays a much more important role than input feature maps in dynamic filer generation. It reveals that annotator encoding is a critical prior that tells the model with which annotator it is dealing and what kind of preferred segmentation results should it produce. Since AVAP adopts two-step training, there is no difference in ’Mean Voting’ performance between these groups of comparative experiments. And experiments show a greater impact on cup segmentation and less impact on the disc.

TABLE V: Ablation study on the effect of different training strategy.

Methods	Average	Mean Voting
Average weight	96.31, 83.36	97.85, 87.54
Weight (0.2, 0.2, 0.6)	96.68, 81.66	97.65, 87.42
Weight (0.4, 0.4, 0.2)	96.59, 83.50	97.78, 86.93
Our (two-step)	96.78, 85.13	97.88, 87.90

3) Training strategy: We adopt a two-step training strategy. Another method that comes to mind directly is to train the model by weighting the loss functions corresponding to the two parts of the AVM and APM into one loss function. Here, we compare the performance of the models under several different sets of weights. As shown in Table V, the two-step approach achieves the best results. Different training strategies have a greater impact on ’Average’ metric. And experiments show a greater impact on cup segmentation and less impact on disc.

V. Conclusion

In this paper, we propose a modeling annotator variation and annotator preference (AVAP) framework for multiple annotations medical image segmentation. It consists of three parts: backbone network, AVM, and APM. AVAP not only produces calibrated image segmentation (output of AVM), but also mimics each annotator and segments medical images with annotator preference (output of APM). AVM model multi-annotations as a multi-class segmentation problem. And we used this method to achieve the first best result on the MICCAI QUBIQ 2021 challenge leaderboards. Further, we are the first to introduce dynamic filter learning into the annotator preference modeling. In this design, compared with the multi-head network, the complexity increase caused by the number of annotators can be ignored. The experiment on the RIGA benchmark with multiple annotations shows that our AVAP framework outperforms a range of state-of-the-art multiple annotations segmentation methods.

Different from existing methods, AVAP adopts a completely different network structure. The experiments show that the annotations variation among multiple annotators of the disc is small. Our model mainly improves the performance of modeling annotator preference. And AVAP obtains the best-calibrated segmentation results on both disc and cup. Compared with MR-Net and PADL state-of-the-art methods, AVM only adopts a very simple network structure. This further illustrates the

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on March 24,2023 at 06:50:19 UTC from IEEE Xplore. Restrictions apply.

effectiveness of modeling multi-annotations as a multi-class segmentation problem. In future work, we can extend the AVM module to introduce better network structures or loss functions to further improve the performance of AVM.

Acknowledgment

This study is supported by grants from the National Natural Science Foundation of China (62106113, 62276081 and 62106115), the Innovation Team and Talents Cultivation Program of National Administration of Traditional Chinese Medicine (NO:ZYYCXTD-C-202004), Basic Research Foundation of Shenzhen Science and Technology Stable Support Program (GXWD20201230155427003-20200822115709001), and The Major Key Project of PCL. Thanks for the support provided by the OpenI Community [https://git.openi.org.cn].

References

[1] Despotović I, Goossens B, Philips W. MRI segmentation of the human brain: challenges, methods, and applications[J]. Computational and mathematical methods in medicine, 2015, 2015.

[2] Hesamian M H, Jia W, He X, et al. Deep learning techniques for medical image segmentation: achievements and challenges[J]. Journal of digital imaging, 2019, 32(4): 582-596.

[3] Zlateski A, Jaroensri R, Sharma P, et al. On the importance of label quality for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1479-1487.

[4] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.

[5] Tang Y, Yang D, Li W, et al. Self-supervised pre-training of swin transformers for 3d medical image analysis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 20730-20740.

[6] Karimi D, Dou H, Warfield S K, et al. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis[J]. Medical Image Analysis, 2020, 65: 101759.

[7] P. Suetens, “Fundamentals of medical imaging, 3rd edition,” 2017.

[8] Mehta R, Filos A, Baid U, et al. QU-BraTS: MICCAI BraTS 2020 Challenge on Quantifying Uncertainty in Brain Tumor Segmentation–Analysis of Ranking Metrics and Benchmarking Results[J]. arXiv preprint arXiv:2112.10074, 2021.

[9] Kofler F, Ezhov I, Fidon L, et al. Deep Quality Estimation: Creating Surrogate Models for Human Quality Ratings[J]. arXiv preprint arXiv:2205.10355, 2022.

[10] Rupprecht C, Laina I, DiPietro R, et al. Learning in an uncertain world: Representing ambiguity through multiple hypotheses[C]//Proceedings of the IEEE international conference on computer vision. 2017: 3591-3600.

[11] Fu M C, Buerba R A, Long III W D, et al. Interrater and intrarater agreements of magnetic resonance imaging findings in the lumbar spine: significant variability across degenerative conditions[J]. The Spine Journal, 2014, 14(10): 2442-2448.

[12] Budd S, Robinson E C, Kainz B. A survey on active learning and human-in-the-loop deep learning for medical image analysis[J]. Medical Image Analysis, 2021, 71: 102062.

[13] Minaee S, Boykov Y Y, Porikli F, et al. Image segmentation using deep learning: A survey[J]. IEEE transactions on pattern analysis and machine intelligence, 2021.

[14] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[15] Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 764-773.

[16] Almazroa A, Alodhayb S, Osman E, et al. Agreement among ophthalmologists in marking the optic disc and optic cup in fundus images[J]. International ophthalmology, 2017, 37(3): 701-717.

[17] Guan M, Gulshan V, Dai A, et al. Who said what: Modeling individual labelers improves classification[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1).

[18] Liu Q, Dou Q, Yu L, et al. MS-Net: multi-site network for improving prostate segmentation with heterogeneous MRI data[J]. IEEE transactions on medical imaging, 2020, 39(9): 2713-2724.

[19] Jensen M H, Jørgensen D R, Jalaboi R, et al. Improving uncertainty estimation in convolutional neural networks using inter-rater agreement[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2019: 540-548.

[20] Chen G, Xiang D, Zhang B, et al. Automatic pathological lung segmentation in low-dose CT image using eigenspace sparse shape composition[J]. IEEE transactions on medical imaging, 2019, 38(7): 1736-1749.

[21] Quantification of Uncertainties in Biomedical Image Quantification Challenge 2021. https://qubiq21.grand-challenge.org/. Accessed 11 Aug 2021

[22] Chen Y, Dai X, Liu M, et al. Dynamic convolution: Attention over convolution kernels[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11030-11039.

[23] Oktay O, Schlemper J, Folgoc L L, et al. Attention u-net: Learning where to look for the pancreas[J]. arXiv preprint arXiv:1804.03999, 2018.

[24] Gu Z, Cheng J, Fu H, et al. Ce-net: Context encoder network for 2d medical image segmentation[J]. IEEE transactions on medical imaging, 2019, 38(10): 2281-2292.

[25] Zhou Z, Siddiquee M M R, Tajbakhsh N, et al. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation[J]. IEEE transactions on medical imaging, 2019, 39(6): 1856-1867.

[26] Jiang Y, Wronski B, Mildenhall B, et al. Fast and High-Quality Image Denoising via Malleable Convolutions[J]. arXiv preprint arXiv:2201.00392, 2022.

[27] Mehrtash A, Wells W M, Tempany C M, et al. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation[J]. IEEE transactions on medical imaging, 2020, 39(12): 3868-3878.

[28] Ji W, Yu S, Wu J, et al. Learning calibrated medical image segmentation via multi-rater agreement modeling[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 12341-12351.

[29] Mirikharaji Z, Yan Y, Hamarneh G. Learning to segment skin lesions from noisy annotations[M]//Domain adaptation and representation transfer and medical image learning with less labels and imperfect data. Springer, Cham, 2019: 207-215.

[30] Liao Z, Hu S, Xie Y, et al. Modeling Annotator Preference and Stochastic Annotation Error for Medical Image Segmentation[J]. arXiv preprint arXiv:2111.13410, 2021.

[31] Yu S, Zhou H Y, Ma K, et al. Difficulty-aware glaucoma classification with multi-rater consensus modeling[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2020: 741-750.

[32] Jia X, De Brabandere B, Tuytelaars T, et al. Dynamic filter networks[J]. Advances in neural information processing systems, 2016, 29.

[33] Zhang J, Xie Y, Xia Y, et al. DoDNet: Learning to segment multi-organ and tumors from multiple partially labeled datasets[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 1195-1204.

[34] Li G, Li C, Zeng C, et al. Region focus network for joint optic disc and cup segmentation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(01): 751-758.

[35] Tian Z, Shen C, Chen H. Conditional convolutions for instance segmentation[C]//European conference on computer vision. Springer, Cham, 2020: 282-298.

[36] Žukovec M, Dular L, Špiclin Ž. Modeling Multi-annotator Uncertainty as Multi-class Segmentation Problem[C]//International MICCAI Brainlesion Workshop. Springer, Cham, 2022: 112-123.

[37] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[38] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.

[39] Zhang L, Tanno R, Xu M C, et al. Disentangling human error from ground truth in segmentation of medical images[J]. Advances in Neural Information Processing Systems, 2020, 33: 15750-15762.

[40] Hendrycks D, Lee K, Mazeika M. Using pre-training can improve model robustness and uncertainty[C]//International Conference on Machine Learning. PMLR, 2019: 2712-2721.

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on March 24,2023 at 06:50:19 UTC from IEEE Xplore. Restrictions apply.